Naive Bayes

① Probability

② Baye's Theorem

Independent Events

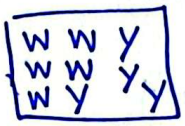
Rolling a dice

$\{1,2,3,4,5,6\}$

$P(1) = 1/6,\ P(2) = 1/6,\ P(3) = 1/6$

Dependent Events

What is the probability of removing a white marble then a yellow marble.



$P(W) = 5/9,\ P(Y) = 4/8$

$P(A \text{ and } B) = P(A)\,P(B/A)$

$P(W \text{ and } Y) = P(W)\,P(Y/W)$

$P(Y/W)$: conditional probability

Bayes Theorem

$P(A \text{ and } B) = P(B \text{ and } A)$

$P(A)\,P(B/A) = P(B)\,P(A/B)$

$$P(A/B) = \frac{P(A)\,P(B/A)}{P(B)}$$

↓

Bayes Theorem

x_1	x_2	x_3	y
-	-	-	Y
-	-	-	N
-	-	-	N

$$P(Y/(x_1,x_2,x_3)) = \frac{P(Y)\,P((x_1,x_2,x_3)/Y)}{P(x_1,x_2,x_3)}$$

$$P(Yes/(x_1, x_2, x_3)) = \frac{P(Yes)\,P(x_1/Yes)\,P(x_2/Yes)\,P(x_3/Yes)}{\cancel{P(x_1)P(x_2)P(x_3)}} \Rightarrow \text{constant}$$

$$P(No/(x_1, x_2, x_3)) = \frac{P(No)\,P(x_1/No)\,P(x_2/No)\,P(x_3/No)}{\cancel{P(x_1)P(x_2)P(x_3)}} \Rightarrow \text{constant}$$

$$[Yes, No] = [0.60, 0.40]$$

⟶ Example

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$P(Yes) = 9/14$

$P(No) = 5/14$

$P(Sunny/Yes) = 2/9$

$P(Overcast/Yes) = 4/9$

$P(Rain/Yes) = 3/9$

$P(Sunny/No) = 3/5$

$P(Overcast/No) = 0$

$P(Rain/No) = 2/5$

$P(Hot/Yes) = 2/9$

$P(Hot/No) = 2/5$

$P(Mild/Yes) = 4/9$

$P(Mild/No) = 2/5$

$P(Cool/Yes) = 3/9$

$P(Cool/No) = 1/5$

$P(High/Yes) = 3/9$

$P(High/No) = 4/5$

$P(Normal/Yes) = 6/9$

$P(Normal/No) = 1/5$

$P(Strong/Yes) = 3/9$

$P(Strong/No) = 3/5$

$P(Weak/Yes) = 6/9$

$P(Weak/No) = 2/5$

Outlook

	Yes	No
Sunny	2	3
Overcast	4	0
Rain	3	2

Temperature

	Yes	No
Hot	2	2
Mild	4	2
Cool	3	1

Humidity

	Yes	No
High	3	4
Normal	6	1

Wind

	Yes	No
Strong	3	3
Weak	6	2

Test Data

Sunny, Hot
High, Strong

$P(Yes)\,P(Sunny/Yes)\,P(Hot/Yes)\,P(High/Yes)\,P(Strong/Yes)$

$= (9/14)(2/9)(2/9)(3/9)(3/9)$

$= 0.0035 \Rightarrow \frac{0.0035}{0.0035 + 0.041} = 0.08$

$P(No)\,P(Sunny/No)\,P(Hot/No)\,P(High/No)\,P(Strong/No)$

$= (5/14)(3/5)(2/5)(4/5)(3/5)$

$= 0.041 \Rightarrow \frac{0.041}{0.0035 + 0.041} = \boxed{0.92}$

→ Naive Bayes Variants

> Bernoulli Naive Bayes

Whenever features are following a Bernoulli distribution [0,1] or [Pass, Fail]

f_1	f_2	output
Yes	Male	Yes
Yes	Female	No
No	Female	Yes
No	Male	Yes

> Multinomial Naive Bayes

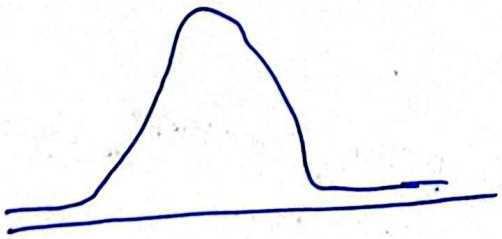
Whenever input data is in form of text

Dataset

Email Body	Spam/Not Spam
Million Dollars	Spam
Promotion	Not Spam

> Gaussian Naive Bayes

If the features are following Gaussian Distribution, and have continuous features



Examples

① IRIS Dataset

Age	weight	height	output
25	170	78	Yes
38	180	75	Yes
22	150	60	No
24	170	.35	Yes